计算机科学与探索
Journal of Frontiers of Computer Science and Technology
ISSN 1673-9418,CN 11-5602/TP

《计算机科学与探索》网络首发论文

题目：医疗健康大数据隐私保护综述
作者：郭子菁，罗玉川，蔡志平，郑腾飞
网络首发日期：2020-11-06
引用格式：郭子菁，罗玉川，蔡志平，郑腾飞．医疗健康大数据隐私保护综述．计算机科学与探索. https://kns.cnki.net/kcms/detail/11.5602.TP.20201105.1508.024.html

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2020-11-06 13:29:01
网络首发地址：https://kns.cnki.net/kcms/detail/11.5602.TP.20201105.1508.024.html

计算机科学与探索
Journal of Frontiers of Computer Science and Technology doi: 10.3778/j.issn.1673-9418.2009071

医疗健康大数据隐私保护综述

郭子菁，罗玉川+，蔡志平，郑腾飞
国防科技大学 计算机学院，长沙 410073
+ 通信作者 E-mail: luoyuchuan09@nudt.edu.cn

摘 要：随着智能移动设备普及化、医疗设备数字化及电子病历结构化的推进，医疗数据呈现爆发增长的特点。在深入研究探讨医疗大数据发展规律，提高对医疗大数据真实价值的认识的同时，如何有效保护数据的隐私安全现已成为广受关注的重要议题。医疗大数据自身特点以及存储环境等都为隐私保护带来了不小的挑战。本文的主要工作为：首先，介绍了医疗大数据的相关概念以及特点。然后，围绕医疗大数据生命周期的四个阶段：数据的采集、存储、共享以及分析，分别介绍面临的风险挑战以及相应的隐私保护技术，并对不同技术的优缺点、适用范围等进行分析。在数据采集时，匿名技术、差分隐私可以抵御数据集成融合带来的基于背景知识的攻击。在存储阶段，医疗大数据多存储于云平台，为了数据的机密性和完整性，常使用加密、审计的方法。在数据共享阶段，主要使用访问控制方法来控制获取数据的对象。在数据分析阶段，在机器学习框架下对医疗健康大数据进行隐私保护以避免暴露敏感数据的隐私风险。最后，针对贯穿医疗大数据生命周期的普遍隐私保护挑战，从管理的层面提出合理的建议。

关键词：医疗大数据；生命周期；隐私保护技术

文献标志码：A **中图分类号：**TP309

郭子菁，罗玉川+，蔡志平，等. 医疗健康大数据隐私保护综述[J]. 计算机科学与探索
GUO Zijing, LUO Yuchuan+, CAI Zhiping, et al. Overview of Privacy Protection Technology of Big Data in healthcare[J]. Journal of Frontiers of Computer Science and Technology

Overview of Privacy Protection Technology of Big Data in healthcare

GUO Zijing, LUO Yuchuan+, CAI Zhiping, ZHENG Tengfei
College of Computer, National University of Defense Technology, Changsha 410073,China

Abstract: With the popularization of smart mobile devices, the digitalization of medical devices and the

*The National Key Research and Development Program of China under Grant Nos. 2020YFC2003400, SQ2019ZD090149 (国家重点研发计划); the National Science and Technology Major Project for IND (investigational new drug) No. 2018ZX09201-014 (国家科技重大专项重大新药创制); the National Nature Science Foundation of China under Grant No. 62072465 (国家自然科学基金) ; the National University of Defense Technology under Grant No. ZK19-38 (国防科技大学科研计划).

structuring of electronic medical records, medical data has shown the characteristics of explosive growth and mass aggregation. It has attracted wide attention to improve the understanding of the real value of big data in healthcare, by doing in-depth research and discussion on its development regulation. However, the issue that how to protect the privacy security effectively in the process deserves our attentions as well. Due to the characteristics of big data in healthcare and the storage environment, privacy protection faces severe challenges. The main work of this paper is as follows: First, the related concepts and characteristics of big data in healthcare are introduced. Then, focusing on the four stages of the life cycle model of big data in healthcare, which includes data collection, storage and share, this paper respectively introduces the risks and challenges it faces and the corresponding privacy protection technologies to do with them, analyzing the merits, drawbacks and their applicable scope. When collecting, anonymous technology and differential privacy can resist attacks based on background knowledge brought by data integration and fusion. In the storage stage, big data in healthcare are mostly stored on the cloud platform. Encryption and auditing are often used for the confidentiality and integrity of data. Talking about the data share stage, the access control plays an important part. During the analysis stage, Privacy Preserving Data Mining methods are adopted to achieve privacy protection is achieved based on the framework of machine learning. Last but not least, regarding to the universal privacy protection challenges throughout the life cycle of big data in healthcare, reasonable suggestions are proposed in the management level.

Key words: big data in healthcare; life cycle; privacy protection technical

1 引言

医学技术与信息技术的不断融合突破，为医疗数据的产生提供了源源不断的动力，也为大数据技术在医疗领域的应用和发展奠定了稳固的基石。医疗数据具有数据量庞大、增长速度快、数据结构多样化和应用价值高等特点，属于大数据的一种。采集、治理及分析这些医疗大数据、有效发掘数据中的潜在价值，在推动临床科研的进步，临床决策支撑以及药物研发等方面都起到了积极的推动作用[1]。因此，健康医疗大数据建设在国内外都受到高度重视，一些发达国家已经搭建了相对成熟的平台，我国由于起步晚而目前专注于数据采集阶段，对于数据的分析处理能力较弱。

然而，在享受从医疗数据中获得有价值的信息为临床科研、健康管理、公共卫生等方面的研究注入新的活力的同时，也不可避免地带来隐私泄露的问题。例如，从 2019 年 7 月中旬到 2019 年 9 月初，Greenbone Networks 分析了全球数千个在线医疗服务系统，发现 2400 多万份来自不同国家的患者数据记录可以在互联网上被访问或轻易下载[2]。泄露的患者数据记录中包含着详细的个人和医疗细节：姓名、出生日期、检查日期、调查项目、主治医师、检测结果的图像信息等。这些数据可被攻击者利用，发布个人姓名和图像以此来损害一个人的声誉；将泄露的数据与其他数据关联起来，从而实现网络钓鱼和社交工程；阅读并自动处理数据来搜索有价值的身份信息，例如利用证件号码用来盗用身份。

如何在不泄露患者隐私的前提下，提高医疗数据的利用率，挖掘其中蕴藏的价值，是目前制约其发展的一个重要因素。因此，在医疗健康大数据的全生命周期中，需要在充分利用数据的同时严密防范隐私泄露，力图在数据利用和隐私保护二者之间找到一个平衡。

1.1 医疗大数据的来源及特征

随着医疗领域信息化的推进，医疗健康方面的电子数据正以前所未有的速度爆发式增长，其类型也多种多样，其中包括患者疾病诊疗数据、身体健康数据以及医疗临床实验数据等。这些类型复杂的规模巨大的医疗数据汇聚起来而呈现出大数据的特性，也就共同构成了医疗大数据。本节主要归纳了医疗大数据的来源以及特点，如图 1 所示。

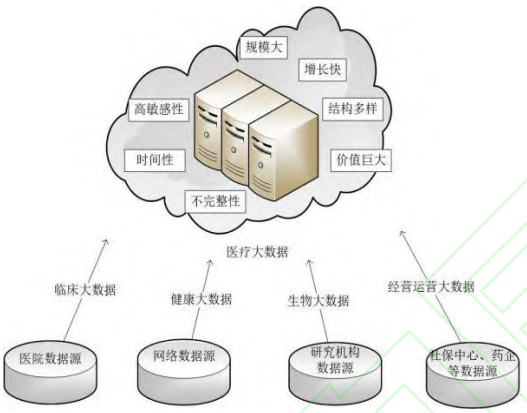

Fig.1 Sources and characteristics of big data in healthcare

图 1 医疗健康大数据来源及特点

医疗健康大数据的来源可被划分为以下四类：

临床大数据：这部分数据主要产生于患者就医过程中，构成了医疗健康大数据的基础内容。患者在就医过程中产生了一系列包含其隐私的数据。首先需提供姓名、年龄、住址、电话等详细的个人信息，在诊疗过程中由医生根据经验判断直接记载或经由各种医疗器械检测产生的电子病历数据、医学图像数据以及使用药物记录等都是临床数据的一部分。此外，在就医过程中还会涉及到相关费用信息、医保使用情况等，这些信息也会被记录下来，在大数据条件下，这些数据经由系统分析，能够产生新的价值。但是，这其中也直接包含着大量个人信息，一旦被非法第三方获取，则直接对患者隐私造成威胁。

健康大数据：随着生活智能化，可穿戴式设备、手机应用渗透到我们的生活中，其获取的信息能帮助每个人监测并记录详细的个人体征数据；在各大网站中浏览、咨询关于疾病、健康等相关内容的行为会暴露出个人偏好数据。这些数据通过互联网与医疗机构相连接，构成电子健康档案内容，用以时刻监控每个人健康情况。这些记录着个体详细健康状况的实时数据，通过网络汇集，就导致了可能暴露健康状况、位置、个人喜好等一系列敏感信息。

生物大数据：得益于高通量测序技术的快速发展，生命科学相关研究机构数据产出能力也日益增强，能够产生包括基因组学、转录组学、蛋白组学、代谢组学等不同组学的庞大数据集。这些生物数据中潜在的巨大价值，不仅有效地推动了生物科研领域的发展，也在农业、健康和医学等领域得以应用。但是，基因检测数据与病理数据相结合时，很容易匹配到具体的个体，在隐私泄露的同时还极易引起基因歧视而给患者带来双重伤害。

经营运营大数据：在各个医疗机构经营运营过程中，也会相应地产生大量数据，例如，运营的成本核算数据、药品、耗材、器械采购数据，药物研发数据，消费者购买行为数据等。数据中涉及药物或相关器械交易记录也往往暴露了用户的身体状况、财政状况等隐私信息，在隐私保护中也是不可忽视的内容。

医疗大数据符合大数据的共同特征——规模大、增长快、结构多样、价值巨大。此外，医疗大数据还具有其他独有的性质。

高度敏感性：医疗大数据中常常直接记录着病人的详细个人信息以及身体健康状况，相较于其他数据具有更高的敏感性，对隐私保护的要求更高。

不完整性：由于医疗健康数据的采集和处理过程常常无法做到紧密衔接，所以医疗数据库中的数据虽然规模庞大但仍然难以全面记录下所有的疾病信息。此外，由于电子病历尚未全面普及，大量数据来源于人工记录，记录内容的偏差和残缺，言语表达的不确定性，资料保管的不到位，都是医疗健康大数据不完整性的源头。

时间性：患者的就诊、发病过程在时间上有一个进度变化，医疗检测的波形、图像数据等都具有一定的时序性。患者的健康状况不是一成不变的，而是始终处于动态变化中，这也就意味着其敏感属性的对应值在随时间变化。

1.2 全生命周期的医疗大数据隐私保护

对医疗大数据而言，从其采集、存储、共享到分析的过程中，均涉及到多方用户，每一个环节都存在严重的隐私泄露忧患，需要采取相应的技术手段来应对。同时也有一部分隐私问题是始终贯穿于所有环节的，可以通过采取适当的管理措施来解决。本文主要从医疗大数据生命周期的几个环节分别阐述存在的隐私泄露挑战以及相应的隐私保护技术，最后从医疗大数据的管理层面提出一些合理的建议（图 2）。

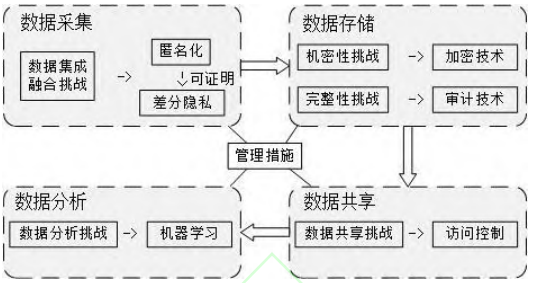

Fig.2 Full life cycle of big data in healthcare privacy-perserving model

图 2 全生命周期医疗大数据隐私保护模型

（1）数据采集

数据采集是医疗健康大数据生命周期中的基础环节。随着信息技术发展，医疗健康渗透到我们生活中方方面面，医疗数据可能来自于医疗机构的信息系统、可穿戴设备、网络等。在数据采集阶段，需要做的就是将各种不同来源的医疗健康大数据汇集在一起，为后续的存储、共享以及分析奠定数据基础。与典型的数据采集不同，医疗数据的采集中直接包含着患者提交的私人信息，即医疗数据具有高度的敏感性。由于医疗健康大数据极其敏感，如何在数据可用的情况下做到高效地隐藏可能泄露用户隐私的内容，是目前亟待解决的问题。

（2）数据存储

数据存储阶段关注的是大规模医疗健康数据的存储管理中的隐私风险。医疗健康大数据因其庞大的数据规模，在采集后需要在云平台进行存储。存储在云平台的数据，其存储者和所有者是完全分离的，而云存储服务提供商并不是完全可信的，因此，存储在云平台的医疗数据并不安全，面临着被不可信的第三方偷窥或者篡改的风险。

（3）数据共享

存储在不同医疗机构的数据通过数据共享，才能达到效益最大化，但在大数据环境下，数据共享带来便利的同时，给患者也带了风险。当患者的数据存储在云平台上，患者并不知道谁访问了共享账户中的数据，因此有很高的数据泄露风险，并且数据泄露后无法追踪，对于隐私保护而言是一个较大的挑战。

（4）数据分析

医疗大数据只有通过分析才能更好地推动疾病诊断、药物研发等医疗领域的发展，也能更好地为患者提供服务。即使经过了匿名化、加密等处理，医疗数据在一系列聚类、关联等数据分析之后，患者的敏感信息仍然有可能会暴露出来。隐私安全不仅需要防止原始数据中的敏感信息泄露，也需要考虑到数据挖掘与分析预测的结果。

2 医疗数据采集中的隐私保护技术

数据采集是数据生命周期中最基础的步骤，医疗健康大数据的采集为科研和机构间的合作提供了便利，但同时也给数据隐私带来了潜在的威胁。

在此阶段存在的风险是基于数据集成融合的链接攻击或其他更复杂的基于知识背景的攻击。患者的诊疗数据、药品或医疗器械的购买记录、互联网上的相关社交信息等医疗数据能够服务于数据分析，同时也一定程度地反映出用户的行为活动。如果攻击者从网络传输中拦截这些数据，并综合利用其他外部信息，从而能够推断出个体身份，这给保护患者隐私带来了严重的挑战[3]。

传统的医疗数据隐私保护主要采用匿名技术，最根本的思想是隐藏数据与个体之间的联系，但简单的删除数据中的个体属性极易通过链接攻击来破解[4]。应对这一攻击手段，k-anonymity[5]被提出，其理念是让数据中的准标识符（不可辨认的属性可对应多个个体，例如，出生日期和邮政编码）可以匹配至少 k 个个体，这意味着一个特定的信息不能区别其他 k-1 个人信息数据集。为了抵抗基于 k-anonymity 的同质性攻击和背景知识攻击，l-diversity 模型[6]被提出，它在 k-anonymity 的基础上，要求每个敏感属性至少包含 l 个表现良好的值。t-close[7]是 l-diversity 模型的进一步细化，l-diversity 模型通过减少数据表示的粒度来保护隐私，通过考虑属性值的分布来区别对待不同的属性值，这是一种为了获得一些隐私而导致数据挖掘有效性损失的权衡。

但是，现有的匿名技术有一个普遍的缺陷——过分依赖攻击者的背景知识假设，并且对其隐私保护水平无法提供严格有效的证明。差分隐私引入医疗领域就有效地解决了匿名技术存在的这些问题。应用差分隐私保护模型时就不必考虑攻击者已经获取的背景知识，其次，差分隐私提供了严格的数学定义和度量隐私泄露的方法，这个特点使得能够比较使用不同参数进行处理的数据集的可用性程度[4]。

2.1 匿名技术

数据匿名性在一定程度上为数据的隐私性提供了保障，在典型的匿名保护方法 k-anonymity，l-diversity，t-closeness 模型的基础上，一些更适用于医疗大数据的匿名技术被提出。

针对数据规模大的问题，Song 等人

提出随机 k 匿名方法[8]。由于寻找匿名等价的过程非常耗时，因此采用两步聚类的方法将原始数据集划分为等价类。首先将原始数据集分成几个不同的子数据集，然后在子数据集中形成等价类，从而大大降低了寻找匿名等价类的计算代价，并且匿名数据集的信息损失小得多，数据的可用性得到了更好的保障。

收集的医疗健康数据通常多种不同类型的敏感属性，因此，在操作高维度数据时，这些不同类型的敏感属性之间的关联与混合同样值得重视。在这种情况下，(a,k)-匿名隐私保护方法将更加有效[9]。Li 等人以 (a,k)-anonymity 模型作为数据采集的隐私保护方案，提出了一种新的基于匿名的医疗保健服务的数据采集方法[10]，采用客户端-服务器-用户模型进行分析。在客户端，利用 (a,k)-anonymity 的概念来生成匿名元组以抵抗可能的攻击，并采用自下而上的聚类方法来创建满足基本匿名隐私级别的聚类。在服务器端，通过泛化技术降低通信成本，通过基于 upgmaa 的聚类组合方法压缩匿名数据，使数据满足更深层次的隐私级别。

由于医疗大数据具有不完整性，为了避免这一特点带来的信息可用性的降低，裴孟丽在 l-diversity 的基础上提出了匿名算法 DAIMDL[11]。DAIMDL 算法在聚类基础上对数据记录进行分组，优化分组后，对划分好的各数据组进行泛化。聚类阶段，基于信息熵的距离计算进行聚类，保证簇内信息距离最小、簇间信息距离最大；泛化阶段，对划分好的各数据组进行泛化，最后得到每个分组内准标识符属性取值相同的各等价类。病人信息经过 DAIMDL 算法处理可避免数据表中不完整数据记录的丢弃，减少医疗数据的信息损失。同时对医疗数据中的敏感属性进行多样化分布，各等价类分组中不同敏感属性值不少于 l 种，得到的医疗数据集满足 l-diversity 匿名模型的要求[6]。

考虑到医疗大数据的持续更新特性，数据在不断更新、插入和删除，继续沿用静态匿名技术，则无疑会产生新的隐私泄露的可能。常见的隐私保护模型有基于 l-diversity 多样性的针对增量数据集的安全匿名方法，但是它只能解决数据的插入操作。文献[12]提出了 m-invariance 方法，可以针对数据的插入和删除进行动态发布，通过满足 m-invariance 相关规则以外，加入了伪元组的概念，最大程度保护了隐私。同时在数据发布时，还发布了一张辅助表，用来记录插入伪元组的统计信息。Shi 等人进一步考虑到目标具体的准标识属性和敏感属性都会变化的情况（例如疾病痊愈或恶化，身体指标改变等），提出了一种动态更新方案[13]。该方案应用拉普拉斯噪声机制对结果集的敏感属性进行保护，并将准标识属性和敏感属性分别保存，根据它们的权限给接收方不同的结果，找到一个既能保证信息的可用性，又能实现隐私保护的最佳集群。

匿名技术较好地防止了患者的敏感数据泄露，同时保证了数据的真实性，在实际应用中受到广泛关注，但其中还存在改进的空间。隐私性和可用性间的平衡问题，目前的研究主要集中于减少信息损失，如何找到一个合理的平衡点是需要进一步深入研究的问题。目前采用的匿名化方法多为贪婪式算法，执行效率并不高，

因此需要研究高效的匿名化算法以应对日益剧增的超大容量数据的发布问题。度量和评价标准问题目前还没有统一的匿名化技术度量和评价标准，因此需要致力于该项研究，给匿名化技术一种更为客观合理的评价。此外，如何高效实现个性化匿名，如何根据实际应用快速准确地选择数据表的准标识符，如何解决分布式环境下多数据表的匿名化等都是值得深人思考和研究的问题。

2.2 差分隐私技术

差分隐私[14]较匿名化的隐私模型而言，可成功抵御大部分隐私攻击并能提供可证明的隐私保证。它在最大化医疗数据可用性的同时，还保证患者隐私的泄露在预期控制范围内。差分隐私技术在数据集中添加的噪声量由查询函数的敏感度决定，与数据集的大小无关。对于规模庞大的医疗数据，如果能够将查询函数的敏感度控制在较低的范围内，就可以通过添加少量的噪声来达到隐私保护的目的，极大程度上保护了医疗数据可用性。这使差分隐私成为了一种十分有前景的医疗数据隐私保护模型。

差分隐私技术旨在保护数据隐私的条件下，同时也确保数据查询的精确性。Li 等人[15]首先开发了一种启发式分层查询方法，然后提出了一种用于差分隐私的私有分区算法，以减少计算开销和查询错误。差分隐私在医疗领域的研究多集中于电子健康记录和基因数据[16][17]。在[16]中，作者首先对数据进行加密，然后使用差分噪声机制对其进行干扰，从而保护了基因组和分布的临床数据的隐私。此外，他们还致力于整合生物学和床边（i2b2）框架的信息学，并在降低网络开销的同时增强了其隐私性。同样，作者在[17]中也采用了传统的差分隐私保护方法和双向解密方法来保护基因组数据不被任何攻击者攻击。作者提高了 i2b2 框架在电子基因组数据记录中的保密性和执行时间。此外，作者在[18]开发了一种不同的私有聚合策略，该策略聚合了健康设备数据，也为其用户提供了及时的激励。该策略结合了差分隐私、Boneh-Goh-Nissim 加密系统和 Shamir 秘密共享，提高了用户的安全性和隐私性。该模型采用 java 的 JPBC 库开发，保证了计算量的降低。

针对医疗健康大数据的差分隐私应用研究还存在不少发展空间。随着技术发展和使用需求，人体传感器或可穿戴设备的尺寸越来越小。因此，需要轻量级和复杂性更低的差分隐私算法来适应这种设备。差分隐私在医疗健康大数据生命周期中多个环节都能起到不可小觑的作用，对医疗系统来说是一个至关重要的解决方案。

3 医疗大数据存储中的隐私保护技术

医疗大数据因其规模巨大且增长迅速，而主要依托云平台进行存储[19]。但是云服务提供者并不完全可信，进而使与患者密切相关的医疗健康数据面临着被不可信的第三方偷窥甚至篡改的风险。为了应对以上安全问题，主要使用加密存储技术以保证数据即使被偷窥也不泄露其中蕴含的信息，使用审计技术来验证数据完整性，以确保数据不被篡改。

3.1 保护机密性的加密存储技术

为了保护数据的机密性，必须使用适当的加密方案。使用传统的对称加密方法

对医疗健康大数据进行加密，虽然在加解密速度上有所保证，但因为医疗大数据存储系统面对着大量用户，也导致了传统的对称加密算法的密钥分发过程过于复杂，所以对称加密并不适用于对医疗健康大数据进行加密。非对称加密方法，其密钥相较易于管理，但对于不断增长的医疗健康大数据而言，计算开销过大，所以也同样不适用。数据加密为数据中的隐私带来了保障的同时，也为用户和云平台带来了不小的计算开销，在一定程度上限制了加密数据的使用以及共享，从而可能导致数据中隐藏价值的浪费。因此，适用于医疗健康大数据和云平台特点的加密方法现已成为存储隐私保护的一个重要研究内容。

Narayan 等人[20]将公钥和私钥结合使用设计出基于属性的加密（ABE）方案。密钥由具有访问权限的第三方管理。该技术通过 PEKS 加密算法允许安全的关键字搜索。数据使用高效的对称密钥加密技术进行加密，并使用基于属性的加密使对称密钥可被授权用户访问。私钥通过安全链接（如 SSL）与用户通信，从而防止窃听者了解有关私钥的任何信息。

为了减小计算复杂度并更好地满足用户的个性化需求，Choe 等[21]提出对患者数据进行选择性加密，以减少计算负担，仅对患者选择的项目应用加密。同时也提出了一些适当的密钥管理所需的特性：患者和医生所持有的密钥数量不应很大；密钥存储简单，消耗空间复杂度低；密钥的更新在时间复杂度上要方便高效；密钥中不应包含任何一方的私人信息；当密钥过期或用户离开组时，应该跟踪并撤销所有的密钥。

Yang 等提出了一种基于症状匹配的跨域动态匿名认证组密钥管理系统(CD-AGKMS)克服了移动设备效率不高、计算量大的局限性[22]。该技术改善了来自不同医疗领域的患者无法相互验证身份并建立安全讨论组的情形，支持建立基于症状匹配的群组。对电子健康系统而言，建立症状相同的患者群聊，共享疾病相关信息具有重要意义。该技术实现了基于症状匹配的患者匿名身份验证：为了建立安全的组密钥，所有参与的患者必须进行匿名身份验证。患者的真实身份不会泄露给组内的其他患者，所有的患者都被证实有相同的症状。一个重要的特征是在认证过程中不会显示症状的明文信息。该技术还能够进行动态患者和组管理：系统提供了时间控制的患者撤销机制。根据估计的治疗时间，为每个患者分配一个有效的时间段，该时间段隐式嵌入到患者的部分秘密密钥中。当有效时间过期时，用户的密钥将被撤销。此外，组密钥管理系统允许患者动态加入或离开组。当成员关系更改以保护新的组会话时，将生成新的组会话密钥。这一方案不需要沉重的双线性配对计算，与其他现有的 GKA 方案相比更具有效性和安全性。

为了更高效更安全地对加密数据进行搜索，使用可搜索对称加密（Searchable symmetric encryption, SSE）[23]，强制对外包加密的数据进行关键字搜索，避免了解密过程，从而在不增加数据泄漏的风险的基础上提高了查询效率。SSE 的中心思想是部署一个隐藏的索引表作为元数据，促进对加密数据的搜索[24]。数据所有者需要基于预处理的消息-关键字对创建索引表。

要执行搜索，用户将提供一个搜索令牌，服务器将使用该令牌通过索引进行搜索。如果找到匹配，则将匹配的加密数据返回给用户。

当客户机希望在所有连接的数据库上执行全局查询时，进一步的挑战将是如何有效地同时在所有独立管理的数据库上执行查询并获得聚合的查询结果。这一问题在现有的算法中未得到有效解决。一种可能的解决方案是让分布在网络中的部分集中服务器收集和聚合并行计算的查询，并将聚合的结果返回给查询者。但是，可能需要在这些服务器上小心地部署强大的安全和恢复机制，以保护它们免受拒绝服务攻击。此外，不同医疗数据在敏感性上的区分对于隐私控制也是至关重要的。一种简单的方法是根据敏感性将记录分割成多个部分，并使用不同的密钥对每个部分进行加密，然而，细粒度的分割会使密钥管理任务复杂化。

3.2 保护完整性的审计技术

医疗数据存储在云服务器中，尽管基于云的系统提供了一些好处，但它也存在各种与安全问题。一旦数据被外包，缺乏控制，或者更准确地说，缺乏数据的所有权会危及数据的完整性。有很多原因会使外包数据的完整性面临风险。云服务，就像任何其他 web 服务一样，必须处理可能损害客户端关键数据的软件和硬件故障。在完整性设计过程中（图 3），CSP 可能会有意地选择覆盖任何数据错误，以用于否认它们。为了节省存储空间，CSP 可能倾向于使用离线方法存储一些很少访问的数据，甚至可能删除这些数据。这些原因导致云用户经常使用一种有效的方式对外包数据执行数据完整性审计。

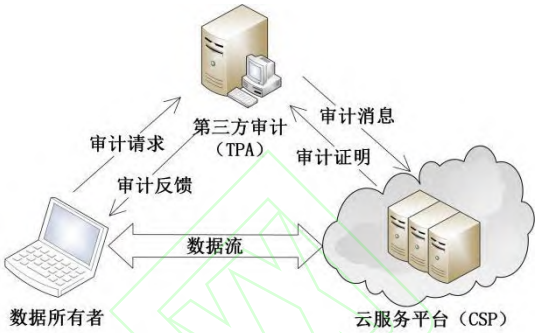

Fig.3 Basic process of integrity audit

图 3 完整性审计基本流程

在访问时检查数据完整性是确保数据拥有的常见方法，但考虑到存储在云上的数据量，在访问时检查数据完整性难以实现。此外，让云提供商或数据所有者审计数据完整性是不恰当的，因为无法保证中立的审计。在这些复杂的、大量的医疗数据存储系统中，数据可能会不时地翻新，为静态数据档案设计的数据审计协议不适合在目前的情况下使用。在这个场景中，需要专门的审计服务定期审计云中的数据完整性。近年来，在不需要访问整个数据的情况下，通过远程服务器检测数据的完整性引起了研究者的广泛关注。

Wang 等人提出一种公开审计方案[25]，主张由第三方审计（third party auditor，TPA）高效地完成数据审计而不增加用户的负担。大多数现有的公开可验证协议，如[26][27][28]支持使用 TPA 对数据进行动态操作。为了避免证书管理问题，人们提出了一些基于身份签名的数据完整性审计协议[29]。这些协议的性能在表 1 中进行了较为详细的罗列和比较。

Table 1　Comparison of partial audit protocol performance
表 1　部分审计协议性能对比

协议	隐私保护	公开可审核性	数据动态	无限查询	批量审核
NaEPASC[26]	×	√	×	√	×
RITS-MHT[27]	×	√	√	√	-
[28]	×	√	√	√	√
[29]	√	√	√	√	-
[30]	√	√	√	√	√
[31]	√	√	×	-	×
Ipad[32]	×	√	-	-	-
[33]	√	√	√	√	×

Gope 等人认为患者对于自己的信息应该拥有掌控权，主张审计日志应该被患者访问和理解[34]。每个患者都应该有权利监控自己的审计数据，并明确谁访问了自己的信息，访问了哪些信息，访问持续了多长时间，访问的目的是什么。患者应该拥有与记录创建、记录如何使用的具体实例、记录更新并最终删除的过程或相关的信息。

然而，审计跟踪只是一种治标不治本的措施，因为在采取应对措施之前，数据的完整性可能已经遭到破坏。但许多系统依赖日志数据的审计作为一种安全机制，当涉及较为严重的问题，如权限滥用、非法访问尝试和患者健康数据的不恰当披露时，审计跟踪可以作为证据。现有的完整性审计技术中，获得用户授权的第三方审计者才能向云服务提供商发起完整性审计挑战，在一定程度上提高了系统的安全性。但针对不同的云类型（如，公有云、私有云、混合云）下的需求，应提出更有效的验证策略。验证效率的进一步优化也是未来的一个研究方向，更高效实时的动态完整性验证方案将为医疗云提供更好的管理服务。

4 医疗大数据共享中的隐私保护技术

每个用户的医疗数据可能存储在不同医院的系统中，也可能保存在使用的智能手机中，而在医疗大数据背景下，这些蕴含巨大价值的数据必然走向共享、开放。比如，分级诊疗、远程医疗、健康管理等新业态的产生，必然驱动数据的有序流动、合理利用和安全分享。

目前已有医疗数据共享平台成功搭建，如美国的 NHIN[35]，不同的医疗机构将患者的检查结果、诊疗记录以及药物使用情况等医疗健康数据通过这个平台进行共享。数据共享带来便利的同时，也不可避免地给患者的隐私也带了安全隐患。

针对这些问题，近年来提出了一些基于访问控制的技术，对这些风险进行了有效地防控。访问控制技术主要通过给不同的用户分配不同的资源访问权限来确保数据仅被某些有权限的特定用户访问。

访问控制技术主要使用两种身份验证：用户身份验证和数据身份验证。用户身份验证可以定义为用户证明其真实性的方式，例如最常见的用户名或带有相关

密码的身份(ID)验证机制[36]。用于确保数据源起源的过程是数据身份验证，最常用的数据认证方法是数字签名方案。

内部和外部攻击者都可以很容易地访问存储在云服务器中的数据，并发起潜在的攻击。应对这一问题，Shamir 和 Tauman 开发了一种称为 hashi -sign-switch 的新范式，它可以将任何签名方案转换为更有效的在线/离线签名方案[36]。Chen 等人解决了上述设计中的关键数据暴露问题[37]。但是他们方案中 trapdoor 哈希函数的散列密钥(HK)是受公钥证书保护的公钥的一部分，不能应用于 ABS 系统，因为签名者是匿名的，他们的公钥是与属性相关的公共参数。因此，在线/离线 ABS 的通用设计仍然是一个开放的问题。

Liu 等人为电子健康系统设计了一个高效、安全的匿名数据认证机制[38]。该机制使用一种应用离散对数的哈希函数来设计 OOABS 的通用方法。该设计不仅可以保护签名者的隐私，保证签名者的匿名性，而且可以防止攻击者伪造签名。Wu 等人实例化了基于安全 ABS 方案的通用结构，并设计了一种分布式验证方案。在他们所设计的系统中，患者可以在移动设备上对数据进行签名，医生和用户可以在不知道签名者的任何属性或身份信息的情况下，对签名者的签名完整性和真实性进行验证。

Zhang 和 Liu 提倡在医疗云中使用匿名数字证书[39]。通过群签名的签名方案，允许一组成员匿名签署电子病历。当参与某个病人会诊的医生对其下一步的治疗得出医学结论时，他们会使用适当的签名算法签署相应电子病历的医学证明。证书将与相应的电子病历一起单独发送给病人。患者可以通过使用该医疗证书和执业者的数字签名来验证咨询结果的真实性。考虑到尊重执业医生的隐私，病人不需要知道签名的执业医生群体，如果在以后出现争议，也可以打开签名以显示签署咨询结果的从业者的身份。同样需要重视的是通信的安全性，目前用于保护在公共网络中传输的信息的技术已经得到了很好的开发和部署，如安全套接字层(SSL)、传输层安全(TLS)、Internet 协议安全(IPSec)等。

针对无线医疗传感网络（wireless medical sensor networks，WMSN）中的安全问题，Kumar 等人提出了一种身份验证协议来监测患者的健康状况，并指出该协议可以抵御已知的安全威胁[40]。但是，He 等人在[41]中提出的工作说明了该协议[40]对于一些安全威胁的抵御是很弱的，He 等人还提出了一种增强的协议，以提高对已知攻击的效率和鲁棒性。Li 等人进一步证明了[41]协议无法检测错误输入，即在登录阶段和密码更改阶段错误输入。Li 等人[42]和 Wu 等人[43]分别提出了使用改进的使用智能卡和哈希函数的用户认证协议来消除[41]协议的漏洞，提供了一个通过不安全的网络远程监控病人健康状况的平台。在现有的认证协议中，研究人员将用户匿名性、用户不可跟踪性、相互认证、对不同攻击的攻击弹性、传感器节点的能量消耗等作为适合医疗技术应用的认证协议的关键因素。Amin 等人[44]在 WMSN 中设计了一个更健壮、更人性化的患者监护系统。提出了一种降低了传感器节点的能耗的健康监测系统体系结构，基于哈希函数的互认证和会话密钥协商协议，为医疗专业人员提供了用户匿名性。经过一系列验

证，该协议在 OFMC 和 CL-AtSe 模型中对主动攻击和被动攻击都是安全的，比同类现有协议具有更强的鲁棒性和安全性。

近年来，人们提出了一些适用于远程医疗信息系统的基于智能卡的密码认证（双因素认证）方案。Xiong 等人[45]使用 Chaudhry 等人的方案作为案例研究，证明了远程医疗信息系统中双因素认证方案对离线字典攻击是不安全的，并且被盗或丢失的智能卡无法撤销。在此基础上，Xiong 提出了一种改进的双因素匿名认证方案。利用随机 oracle 模型和 Burrows Abadi Needham 逻辑给出了该方案的安全性分析。

大数据环境以及无线移动网络环境为医疗大数据的访问控制带来了诸多挑战。随着计算能力的进一步提升，访问控制的效率得到快速提升。同时，巨大的数据量用于身份验证，从而可以实现更加精准、更加个性化的访问控制。目前针对医疗大数据的访问控制研究还在进一步深入，更加尊重用户意愿的细粒度权限分配将会成为重点研究方向。

5 医疗大数据分析中的隐私保护技术

医疗数据的积累、电子病历的推广为机器学习应用于医疗领域奠定了良好的数据基础。医疗大数据只有经过分析处理，才能将其中对于疾病的诊断、治疗和医学研究方面有价值的知识和规则挖掘出来。但是，有些数据表面上并无联系，而通过数据挖掘技术，一些敏感的信息就可能被挖掘出来：独立出现时并不涉及到个人隐私的数据，可能通过和个人信息的匹配后，足以分析出个人敏感信息。

对医疗数据进行数据挖掘，一些原本无法被识别的信息和模式可能会暴露出来并泄露给不可信的第三方。因此需要在保护隐私前提下对数据进行分析处理，限制对大数据中敏感知识的挖掘。虽然医疗大数据经过了一系列清洗操作，使病人的相关隐私无法从数据集中直接得到，但对大量汇集的信息进行挖掘后，一些敏感信息可能会通过挖掘的结果泄露。因此，将机器学习运用于医疗领域的过程中如何进行隐私保护是医疗健康大数据分析方面值得研究的问题。

（1）机密计算

机密计算强调在机器学习的训练过程中对数据进行传输以及计算的机密性，为数据提供隐私保护。当前实现机密计算的方法有可信执行环境，同态加密和多方安全计算。

可信执行环境以硬件安全为强制保障，在计算芯片上独立出一块绝对安全区域，用以保障运行的数据和代码（图 4）。基于可信执行环境——因特尔的 SGX 技术，一种罕见病基因数据分析系统 PRINCESS 被提出[46]，在加密数据上执行安全的分布式计算，并针对川崎病进行了基于家庭的等位基因关联研究，PRINCESS 算法比同态加密和乱码电路等替代方案能够更快地提供安全和准确的分析。Feng C 等人还提出了基于 SGX 技术的基因数据分析框架 PRESAGE[47]，以及新颖的安全的基因亲缘关系分析方法 PREMIX[48]。但是基于硬件安全的 SGX 会遭受到特定算法的旁路攻击[49]。

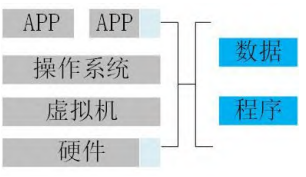

Fig.4 Trusted Executive Environment

图 4 可信执行环境

同态加密，可以在不提供密钥的情况下对密文进行机密的计算，只有使用密钥才能将其解密成明文。在对基因数据进行分析时，考虑到其敏感性，通常应用同态加密技术。例如，基于基因数据的罕见病研究框架 HEALER[50]，在保护人类基因组数据的基础上，分析小样本量的罕见变异体。在一般基因数据分析[51]中，基因组数据所有者只提供加密的序列，公共商业云可以执行序列分析而无需解密，结果只能由数据所有者或持有解密密钥的指定代表解密。在全基因组关联分析计划中[52]，所有基因型和表型数据都进行完全同态加密，允许云对加密的数据执行有意义的计算。但同态加密的实际应用受限于巨大的计算开销，现在的技术大约只能扩展到 MNIST 和 CIFAR 的推断部分[53]。

多方安全计算是参与方以各自隐私数据为输入共同计算一个函数值，各参与方无法获得其他人的隐私数据，只能获得计算结果。由于无需依赖可信任的第三方，安全多方计算技术被广泛应用于生物医疗数据研究中。例如，多机构医疗健康记录匹配算法[54]以及全基因组关联分析算法[55]等。但在实际应用中，节点之间的通信量不容小觑，如何减小这一通信开销，也是当下的一个研究热点[56]。

（2）模型隐私

训练后的模型也可能会造成训练数据的隐私泄露。因为机器学习的模型都会记住自己的训练数据，从而导致发布模型会有训练数据隐私泄露的风险。

而差分隐私可以衡量和控制模型对训练数据的泄露，刻画出单个数据样本对模型的影响。差分隐私技术机器学习算法结合，可确保健康数据的完全隐私（如表 2）。

Table 2　The combination of differential privacy technology and machine learning algorithm

表 2　差分隐私技术与机器学习算法结合的场景

	隐私机制	使用的差分隐私方法	优点	隐私标准	使用平台
[57]	医疗健康数据库的隐私和安全管理	利用差分隐私的拉普拉斯噪声来增强数据隐私	轻量级框架，支持复杂的数据挖掘任务和各种 SQL 查询；减少计算开销	(ε,Δ)-差分隐私	Prototype in JAVA16 使用大整数
[58]	医疗数据范围查询差分隐私算法	利用拉普拉斯噪声实现了数据分区和工作负载	优化查询错误率	ε-差分隐私	N/A
[59]	端到端差分隐私的深度学习方法	基于差分隐私随机梯度下降的深度学习方法	提高训练的准确性和效率	(ε,δ)-差分隐私	N/A
[60]	医疗数据差分隐私数据聚类框架	基于差分隐私机器学习的 k-means 聚类	优化隐私分配预算 提高学习精度	(ε,δ)-差分隐私	Hadoop

2019 年提出的高斯差分隐私[61]在计算复合和采样两种操作的隐私损失都给出了一个紧估计，在隐私损失的统计上都更加精准，从而在相同隐私预算下的噪音更小，取得的性能更好。

另一个模型隐私的研究热点是模型遗忘（machine unlearning），即如何让个人控制他们的数据何时可以使用，何时不能使用，也就是“被遗忘权”[62]。实现模型遗忘的最直接方法是在数据集中删除指定的数据后重新训练，但重新训练的计算开销非常高，因此，需要探索的是如何耗费尽可能少的计算开销实现模型遗忘。一种方法是在需要删除数据时对已经训练好的模型作进一步处理，使其与重新训练的模型在统计意义上近似不可区分[63][64]；而另一种方法是设计新的训练方法，降低重新训练的代价，例如在最初训练的时候就将数据分块，每块数据单独训练出子模型，然后汇总子模型的结果，当需要删除数据时只需要重新训练一个子模型，这样就能在一定程度上减少训练成本[65][66]。

（3）联邦学习

联邦学习本质上是一种分布式学习框架。多个医疗机构的数据集中整合训练往往能取得比使用一家机构数据单独训练的效果好，但是每个医疗机构都希望自己的数据是安全的，对数据集中整合往往带来复杂的隐私和数据安全等问题。而通过联邦学习，数据拥有者在不用直接提供数据的情况下，也可得到训练模型，并且模型的训练效果也能得到保证，与数据整合之后的训练效果相差无几。联邦学习技术通过参数交换方式对医疗健康数据进行了有效的隐私保护，数据和模型保留在本地，本身不会进行传输，因此在数据层面不存在泄露的可能。

Kim Y 等人在保证各医院的数据不离开本地的情况下，将多家医院的数据联合分析出特定患者人群的表型[67]。从研究结果可知，单独使用一家医院的数据与联合利用两家医院的数据分析得出的结果差异较大，而使用联邦学习的方式，在数据不出医院的情况下，在准确性和表型发现方面与集中式训练模型相似，同时又尊重隐私。

Brisimi 等人提出了一种联邦优化方案(cPDS)[68]，可用于求解支持向量机问题。他们使用了波士顿医疗中心的心脏记录电子数据集，利用 cPDS 来区分在目标年内患者是否可能住院，并取得了较好的结果。cPDS 框架是通用的，它的优点在于可伸缩性，以及避免了数据交换，这在医疗领域是非常重要的。

NVIDIA 团队在 BraTS 数据集上应用并评估了用于脑肿瘤分割的联邦学习系统[69]。这是第一个用于医学图像分析的隐私保护联邦学习系统，并且探讨了在联邦学习系统中应用差分隐私技术来保护病人数据的可行性。虽然联邦学习可以保证极高的隐私安全性，但通过模型反演，仍可以设法使数据重现。为了进一步提高联盟学习的安全性，研究人员研究了使用 ε-差分隐私框架的可行性。这个框架是一种正式定义隐私损失的方法，可以借助其强大的隐私保障性来保护患者与机构数据。

联邦学习的主要优点是数据可以保留在其所有者手中，同时仍然能够对不同所有者的数据进行训练。联邦拓扑是灵活

的或完全分散的，不需要持续的在线可用性，因为培训可以离线进行，结果可以稍后返回。因此，在医疗领域，联邦学习方法无疑已成为使用最广泛的下一代隐私保护技术。然而，联邦学习在具体实现中计算和通信开销较大，也是当下亟待解决的问题。

6 展望

如何在保证对医疗大数据的较高利用率，挖据数据价值的同时，切实保护用户隐私，是目前医疗研究领域的关键问题。本文首先介绍了医疗健康大数据的复杂来源，以及其区别与一般大数据的特殊性质。然后从医疗大数据生命周期出发介绍了每个环节中存在的隐私保护问题，并对隐私保护的技术进行了分类阐述，简要探讨了各种技术的可取性以及局限性，探索了医疗大数据隐私数据护技术进一步发展的方向。总体而言，在医疗大数据领域，更多的文献提出了相关问题和建议，而真正将技术应用到实践中的较少，隐私保护的细粒度、个性化需求也越来越迫切，将成为今后研究的重点内容。

在医疗健康大数据的生命周期中，采用隐私保护技术能够在一定程度上防止隐私的泄露。但是如果没有科学合理的管理措施，仍会面临人工操作不当、恶意的内部人员、基础设施被破坏、相关法律法规不明确等技术方面难以控制的问题。

（1）建立隐私安全规范与管理标准，完善法律法规

在医疗健康大数据的全生命周期中，离不开工作人员的操作管理，例如医疗部门的医生，能够直接接触到患者的私人信息以及检测结果，这些不仅暴露了患者的身体情况，还透露其家庭住址、生活习惯等信息。为了防止患者敏感信息被恶意使用和泄露，应该制定严格的管理标准，对各个环节中涉及的工作人员进行隐私安全规范培训，并切实落实到其操作管理之中。

（2）完善医疗健康大数据隐私保护法律法规

法律具有强制性，是保障患者隐私，减少数据泄露的有力武器。政府应加快针对医疗健康大数据隐私保护的立法工作，并进一步完善保护制度，对恶意窃取数据的行为加大打击力度。此外，考虑到医疗健康大数据的传输是全球范围内的，建立并完善一套关于医疗健康大数据保护的国际标准法律也十分重要。

（3）基础设施实时监管

医疗健康大数据的隐私安全也依赖于生命周期中各个基础设施的安全，例如存储了医疗数据的云平台，一旦损坏或被恶意攻击，数据可能会丢失、篡改。在医疗健康大数据的全生命周期中，涉及到多种基础设施，每一个环节的隐私安全都不容小觑，需要进行实时监管保护，在第一时间应对突发状况。

References:

[1] Obermeyer Z, Emanuel E J. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine.[J]. N Engl J Med, 2016, 375(13): 1216-1219.

[2] A Review of Cyber Security Incidents in 2019(International)[OL][2020-2-10].https://www.freebuf.com/articles/network/226830.html.

[3] Wang K. A Survey on Risks of Big Data Privacy[C]. International Conference on Applications and Techniques in Cyber Security and Intelligence. Edizioni della Normale, Cham, 2017:

161-167.

[4] Xiong Ping, Zhu Tianqing, Wang Xiaofeng. Differential privacy Protection and application[J]. Journal of Computer Science, 2014, 37(1): 101-122.

熊平, 朱天清, 王晓峰. 差分隐私保护及其应用. 计算机学报, 2014, 37(1):101-122.

[5] Sweeney L. K-anonymity:A model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[6] Machanavajjhala A, Kifer D, Gehrke J. L -diversity: Privacy beyond k-anonymity[J]. Acm Transactions on Knowledge Discovery from Data, 2007, 1(1):3.

[7] Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity[C]//IEEE International Conference on Data Engineering. 2007.

[8] Song F, Ma T, Tian Y, et al. A new method of privacy protection: random k-anonymous[J]. IEEE Access, 2019, PP(99):1-1.

[9] Li H T, Ma J F, Fu S . A privacy-preserving data collection model for digital community[J]. ence China Information ences, 2015, 58(3):1-16.

[10] Li H, Guo F, Zhang W, et al. (a,k)-Anonymous Scheme for Privacy-Preserving Data Collection in IoT-based Healthcare Services Systems[J]. Journal of Medical Systems, 2018, 42(3):56.

[11] Pei Mengli. An anonymous algorithm based on L-Diversity for missing medical data[D]. Henan: Zhengzhou University, 2019.

裴孟丽.基于 l-diversity 面向缺失医疗数据的匿名算法研究[D].河南:郑州大学, 2019.

[12] Xiao X, Tao Y. M-invariance: towards privacy preserving re-publication of dynamic datasets[J]. Proc Sigmod, 2007.

[13] Shi Y, Zhang Z, Chao H C , et al. Data Privacy Protection Based on Micro Aggregation with Dynamic Sensitive Attribute Updating[J]. Sensors, 2018, 18(7).

[14] Dwork C. Differential Privacy[C]//Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II (ICALP'06). 2006, 1-12.

[15] Li H, Dai Y, Lin X. Efficient e-health data release with consistency guarantee under differential privacy[C]//2015 17th International Conference on E-health Networking, Application & Services (HealthCom). IEEE, 2016.

[16] Raisaro J L, Troncoso-Pastoriza J R, Misbach M, et al. MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018, PP(99):1-1.

[17] Louis R J, Gwangbae C, Sylvain P, et al. Protecting Privacy and Security of Genomic Data in i2b2 With Homomorphic Encryption and Differential Privacy[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2017, PP:1-1.

[18] Tang W, Ren J, Deng K, et al. Secure Data Aggregation of Lightweight E-Healthcare IoT Devices With Fair Incentives[J]. IEEE Internet of Things Journal, 2019, 6(5):8714-8726.

[19] Gaff B M, Sussman H E, Geetter J. Privacy and Big Data[J]. Computer, 2014, 47(6):7-9.

[20] Narayan S, Gagné M, Safavi-Naini R. Privacy preserving EHR system using attribute-based infrastructure[J]. Proc ACM workshop on cloud computing security, 2010:47.

[21] Choe J, Yoo S K. Web-based secure access from multiple patient repositories[J]. International Journal of Medical Informatics, 2008,77(4):242 -248.

[22] Yang Y, Zheng X, Liu X, et al. Cross-domain dynamic anonymous authenticated group key management with symptom-matching for e-health social system[J]. Future Generation Computer Systems, 2017:S0167739X1730554X.

[23] Sen Poh G, Chin J J, Yau W C, et al. Searchable Symmetric Encryption: Designs and Challenges [J]. ACM Computing Surveys, 2017, 50(3):1-37.

[24] Li J, Wang Q, Wang C, et al. Fuzzy keyword search over encrypted data in cloud computing[C]//INFOCOM, 2010 Proceedings IEEE. IEEE, 2014.

[25] WANGC, WANGQ, RENK, et al. Privacy- preserving public auditing for data storage security in cloud computing[C]//Proceedings of IEEE INFOCOM, March 15-19, 2010, San Diego, CA, USA. Piscataway: IEEE Press, 2010: 525-533.

[26] Tan S, Jia Y. NaEPASC: a novel and efficient

public auditing scheme for cloud data[J]. Journal of Zhejiang Universityence C, 2014, 15(9):794-804.

[27] Garg N, Bawa S. ITS-MHT: Relative indexed and time stamped Merkle hash tree based data auditing protocol for cloud computing[J]. Journal of network and computer applications, 2017.

[28] Shen J, Shen J, Chen X, et al. An Efficient Public Auditing Protocol With Novel Dynamic Structure for Cloud Data[J]. IEEE Transactions on Information Forensics & Security, 2017, 12(99):2402-2415.

[29] Shang T, Zhang F, Chen X, et al. Identity-Based Dynamic Data Auditing for Big Data Storage[J]. IEEE Trans. Big Data (2019).

[30] Fan Y, Lin X, Tan G, et al. One secure data integrity verification scheme for cloud storage[J]. Future Generation Computer Systems, 2019: 376-385.

[31] Yu Y, Au M, Ateniese G, et al. Identity-based Remote Data Integrity Checking with Perfect Data Privacy Preserving for Cloud Storage[J]. IEEE Transactions on Information Forensics & Security, 2016:1-1.

[32] Zhang J, Li P, Mao J. IPad: ID-based public auditing for the outsourced data in the standard model[J]. Cluster Computing, 2016, 19(1):127-138.

[33] Srinivas J, Das A K, Kumar N, et al. Cloud Centric Authentication for Wearable Healthcare Monitoring System[J]. IEEE Transactions on Dependable & Secure Computing, 2018, PP(99): 1-1.

[34] Gope P, Amin R. A Novel Reference Security Model with the Situation Based Access Policy for Accessing EPHR Data[J]. Journal of Medical Systems, 2016, 40(11):242.

[35] Ma Can. Comparative Research on Sharing of Medical Big data resources at home and abroad[J].Intelligence data work, 2016, 37(3): 63-67.

马灿, MaCan. 国内外医疗大数据资源共享比较研究[J]. 情报资料工作, 2016, 37(3):63-67.

[36] Adi Shamir, Yael Tauman. Improved Online/ Offline Signature Schemes[M]//Advances in Cryptology-CRYPTO 2001. Springer Berlin Heidelberg, 2001.

[37] Chen X, Zhang F, Susilo W, et al. Efficient generic on-line/off-line signatures without key exposure[J]. Applied Cryptography and Network Security. Springer, 2017: 18-30.

[38] Liu J, Ma J, Wu W, et al. Protecting Mobile Health Records in Cloud Computing: A Secure, Efficient, and Anonymous Design[J]. ACM Transactions on Embedded Computing Systems, 2017.

[39] Zhang R, Liu L. Security Models and Requirements for Healthcare Application Clouds[C]// IEEE International Conference on Cloud Computing, CLOUD 2010, Miami, FL, USA, 5-10 July, 2010. IEEE, 2010.

[40] Kumar P, Lee S G, Lee H J. E-SAP: Efficient-Strong Authentication Protocol for Healthcare Applications Using Wireless Medical Sensor Networks[J]. Sensors, 2012, 12(2):1625-1647.

[41] He D, Kumar N, Chen J, et al. Robust anonymous authentication protocol for health-care applications using wireless medical sensor networks[J]. Multimedia Systems, 2015, 21(1): 49-60.

[42] Li X, Niu J, Kumari S, et al. A new authentication protocol for healthcare applications using wireless medical sensor networks with user anonymity[J]. Security & Communication Networks, 2016, 9(15):2643- 2655.

[43] Wu F, Xu L, Kumari S, et al. An improved and anonymous two-factor authentication protocol for health-care applications with wireless medical sensor networks[J]. Multimedia Systems, 2015, 23(2):1-11.

[44] Amin R, Islam S H , Biswas G P , et al. A robust and anonymous patient monitoring system using wireless medical sensor networks[J]. Future Generation Computer Systems, 2016, 80.

[45] Xiong H, Tao J, Yuan C. Enabling Telecare Medical Information Systems with Strong Authentication and Anonymity[J]. IEEE Access, 2017:1-1.

[46] Feng C, Shuang W, Xiaoqian J, et al. PRINCESS: Privacy-protecting Rare disease International Network Collaboration via Encryption through Software guard extensions[J]. Bioinformatics(6): 871.

[47] Chen F, Wang C, Dai W, et al. PRESAGE: Pri-

vacy-preserving genetic testing via Software Guard Extension[J]. Bmc Medical Genomics, 2017, 10(S2):48.

[48] Chen F, Dow M, Ding S, et al. PREMIX: Privacy-preserving Estimation of Individual admixture[J]. AMIA. Annual Symposium proceedings / AMIA Symposium, 2016: 1747-1755.

[49] Wang W, Chen G, Pan X, et al. Leaky Cauldron on the Dark Land: Understanding Memory Side-Channel Hazards in SGX[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017: 2421-2434.

[50] Shuang W, Yuchen Z, Wenrui D, et al. HEALER: homomorphic computation of exact Logistic regression for secure rare disease variants analysis in GWAS[J]. Bioinformatics(2):211.

[51] Cheon J H , Kim M , Lauter K . Homomorphic Computation of Edit Distance[C]//International Conference on Financial Cryptography and Data Security. Springer Berlin Heidelberg, 2015.

[52] Wen-Jie, Lu, Yoshiji, et al. Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption[J]. Bmc Medical Informatics & Decision Making, 2015.

[53] Gilad-Bachrach, Laine, Lauter, et al. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy[C]//Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

[54] Kuzu M, Kantarcioglu M, Inan A, et al.Efficient privacy-aware record integration[J].Proceedings of the 16th International Conference on Extending Database Technology, 2013: 167-178.

[55] Zhang Y, Marina Blanton. Secure distributed genome analysis for GWAS and sequence comparison computation[J]. Bmc Medical Informatics & Decision Making, 2015, 15(Suppl 5):S4.

[56] Wagh S, Gupta D, Chandran N. SecureNN: 3-Party Secure Computation for Neural Network Training[J]. Proceedings on Privacy Enhancing Technologies, 2019, 2019(3):26-49.

[57] Dankar F K, El Emam K. Practicing differential privacy in health care: A review[J]. Trans. Data Privacy, 2013, 6(01):35-67.

[58] Mohammed N, Barouti S, Alhadidi D, et al. Secure and Private Management of Healthcare Databases for Data Mining[C]//IEEE International Symposium on Computer-based Medical Systems. IEEE, 2015.

[59] Alnemari A, Romanowski C J, Raj R K. An Adaptive Differential Privacy Algorithm for Range Queries over Healthcare Data[C]//IEEE International Conference on Healthcare Informatics. IEEE, 2017.

[60] Beaulieu-Jones B K,Yuan W, Finlayson S G, et al. Privacy-Preserving Distributed Deep Learning for Clinical Data[J]. Machine Learning for Health (ML4H) Workshop at NeurIPS, 2018.

[61] Dong J, Aaron Roth, Su J. Gaussian differential privacy[J]. arXiv preprint arXiv: 1905.02383 (2019).

[62] Liu X, Tsaftaris S A. Have you forgotten? A method to assess if machine learning models have forgotten data[C]//23rd International Conference on Medical Image Computing and Computer Assisted Intervention. 2020.

[63] Golatkar A, Achille A, Soatto S. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks[C]//2020 IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[64] Ginart A, Guan M Y, Valiant G, et al. Making AI Forget You: Data Deletion in Machine Learning[J]. Neural Information Processing Systems 32 (NIPS 2019).

[65] Bourtoule, Lucas, et al. Machine Unlearning[J]. Security & Privacy, 2020.

[66] Fredrikson M, Jha S, Ristenpart T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures[C]//the 22nd ACM SIGSAC Conference. ACM, 2015.

[67] Kim Y, Sun J, Yu H, et al. Federated Tensor Factorization for Computational Phenotyping[J]. Kdd, 2017, 2017:887-895.

[68] Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy[J]. Foundations & Trends in Theoretical Computer Science, 2013, 9(3-4): 211-407.

[69] Francis, S, Collins, et al. A new initiative on precision medicine[J]. New England Journal of Medicine, 2015.

GUO Zijing was born in 1997. She is a master candidate at National University of Defense Technology now. Her research interests include information security and privacy protection.
郭子菁（1997-），女，江西省瑞昌市人，国防科大硕士生，主要研究领域为信息安全和隐私保护。

LUO Yuchuan was born in 1990. He is a lecturer at National University of Defense Technology now. His research interests include network and information security.
罗玉川（1990-），男，四川广安人，2019 年获国防科技大学工学博士，现为国防科技大学计算机学院讲师，主要研究领域为网络与信息安全。

CAI Zhiping was born in 1975. He is a professor, PhD supervisor at National University of Defense Technology. Senior member of CCF. His research interests include network security and big data.
蔡志平（1975-），男，湖南益阳人，国防科大计算机学院教授，博士生导师，CCF 高级会员，主要研究领域为网络安全和大数据。

ZHENG Tengfei was born in 1993. He is a Ph.D candidate at National University of Defense Technology now. His research interests include information security and privacy protection.
郑腾飞（1993-），男，河北省石家庄市人，国防科大博士生，主要研究领域为信息安全和隐私保护。